Quantifying and mitigating bias in inference on gravitational wave source populations

Quantifying and mitigating bias in inference on gravitational wave source populations

Jonathan R. Gair[1,*] and Christopher J. Moore[1,†]

[1] *Institute of Astronomy, Madingley Road, Cambridge, CB30HA, United Kingdom*

(Dated: February 12, 2018)

When using incorrect or inaccurate signal models to perform parameter estimation on a gravitational wave signal, biased parameter estimates will in general be obtained. For a single event this bias may be consistent with the posterior, but when considering a population of events this bias becomes evident as a sag below the expected diagonal line of the P-P plot showing the fraction of signals found within a certain significance level versus that significance level. It would be hoped that recently proposed techniques for accounting for model uncertainties in parameter estimation would, to some extent, alleviate this problem. Here we demonstrate that this is indeed the case. We derive an analytic approximation to the P-P plot obtained when using an incorrect signal model to perform parameter estimation. This approximation is valid in the limit of high signal-to-noise ratio and nearly correct waveform models. We show how the P-P plot changes if a Gaussian process likelihood that allows for model errors is used to analyse the data. We demonstrate analytically and using numerical simulations that the bias is always reduced in this way. These results provide a way to quantify bias in inference on populations and demonstrate the importance of utilising methods to mitigate this bias.

I. INTRODUCTION

In the coming years it is expected that the advanced era ground-based gravitational wave (GW) detectors that are now coming online (such as advanced LIGO [1] and advanced Virgo [2]) will begin to make routine measurements of GWs from a variety of sources. Later in this decade, pulsar timing arrays could also begin to detect sources of nanohertz gravitational waves [3–6] and there are ambitious plans for a space-based gravitational wave detector (eLISA [7]) operating in the millihertz band, that with be launched by ESA around 2034. Inferences about source parameters in this new era of GW astronomy will rely on the availability of detailed signal models for the sources. The calculation of accurate models is computationally prohibitive, however, so approximate models will be used for inference, which will, in general, lead to biases in the parameter estimates obtained. This can lead one to make incorrect inferences about individual sources as well as incorrect inferences about astronomical populations of sources. The bias due to incorrect models becomes more important for louder sources. eLISA is expected to observe the inspiral and merger of supermassive black holes at signal-to-noise ratios of $\mathcal{O}(10^3)$. The impact of parameter bias has been shown to be even more significant in this case [8].

A common way to quantify the performance of parameter estimation is via the probability-probability (P-P) plot. The P-P plot shows the probability that the true source parameters will lie in a given confidence interval estimated from the detector data, against the value of the confidence interval. In the ideal, unbiased, case the P-P should be a diagonal line; i.e., $x\%$ of the time the true source parameters should lie with the $x\%$ confidence interval. However, there are a variety of effects that can cause the P-P plot to deviate from this ideal. For example, use of a greedy algorithm to build a multi-dimensional confidence interval from a kD-tree constructed from a random sample of points from a distribution (this problem was discussed in the context of sky-localisation by [9]), deviations between the waveform model and the true signal due to a breakdown of general relativity (GR) in the strong field (the case of undetectable deviations from GR, the so-called "stealth-bias", was considered in [10]), and mis-estimating the noise properties of the detector can all cause the P-P plot to deviate from a ideal diagonal line. However, the cause of biased parameter estimation that we will consider in this paper is the presence of inaccuracies in the waveform model used to analyse the data [8]. If such a systematic error is present the returned confidence intervals from a parameter estimation study will be shifted away from the true parameters making it less likely that the confidence interval contains the true parameters. Therefore the P-P plot will "sag" below the ideal diagonal line.

Recently [11] the authors proposed a marginalised likelihood which uses Gaussian processes (GPs) to fold in extra information from a small *training set* of accurate waveforms, e.g. numerical relativity (NR) waveforms. Accurate here refers to how well these waveforms represent solutions of the GR field equations. Numerical relativity waveforms are not perfectly accurate, but they are the best solutions currently available and inaccuracies in them can be folded into the GP analysis. If astrophysical gravitational waves are governed by a theory other than general relativity, these waveforms will not be accurate representations of reality. This will also lead to a bias, but one that is harder to quantify without knowing the true theory of gravity. Here we proceed assuming GR is correct and look only at biases from model uncertainties. Once observations are made this assumption could

* jrg23@ast.cam.ac.uk
† cjm96@ast.cam.ac.uk

be revisited if evidence arises for departures from GR.

The GP marginalised likelihood in general shifts the best fit parameters closer to the true parameters and broadens the peak in the posterior, making it more likely that a given confidence contour contains the true parameters. Therefore, it would be expected that parameter estimates obtained using the marginalised likelihood would exhibit less of a bias, and the P-P plots would exhibit less of a "sag". However, the Gaussian process regression (GPR) which underlies the marginalised likelihood makes some assumptions about how the error in the waveform model varies over parameter space. In this paper, we investigate the P-P plots both in the case where these assumptions turn out to be correct, and, more importantly, when they are incorrect.

There are two main results in this paper. The first is a derivation of an analytic expression for the expected sag in a P-P plot arising from waveform uncertainties. This is derived under the assumption that the waveform error is small so that we can use the linear signal approximation. The second is that the use of the marginalised likelihood constructed via Gaussian process regression to analyse data leads to a reduction in the size of the deviation from the diagonal line. The sag is removed completely if the true waveform errors are drawn from the same model used to construct the marginalised likelihood. However, even when the errors follow a different distribution, the marginalised likelihood leads to a reduction in the sag.

This paper is organised as follows. Sec. II provides a recap of and quotes some necessary results about GW parameter estimation, and introduces the marginalised likelihood. Sec. III derives analytic expressions for the P-P plots for both the standard and marginalised likelihoods for a variety of possible waveform errors. Sec. IV describes the numerical simulations that were performed to back-up the analytic results in Sec. III. Finally Sec. V contains a discussion of the results and concluding remarks.

II. PARAMETER ESTIMATION

We assume that the source of GWs is fully specified by a parameter vector $\vec{\lambda}$, and that the true waveform model is $h(t;\vec{\lambda})$ (hereafter the dependence of h on time t is supressed for clarity). The aim of a parameter estimation study given measured data s, is to estimate the posterior probability on the parameters, $P(\vec{\lambda}|s)$. This is given from Bayes theorem (Eq. 1) by the likelihood, $P(s|\vec{\lambda}) \equiv L'(\vec{\lambda})$, the prior, $P(\vec{\lambda})$, and the normalising Bayesian evidence $Z = \int \mathrm{d}\vec{\lambda}\, P(\vec{\lambda})L'(\vec{\lambda})$;

$$P(\vec{\lambda}|s) = \frac{P(\vec{\lambda})L'(\vec{\lambda})}{Z}\,. \tag{1}$$

As this paper concerns parameter estimation, and not model selection, we will not discuss the evidence further, since for any given source, this just enters as a normalisation factor for the posterior. In the case of stationary, Gaussian, additive noise n in the detector the measured data is given by $s = h(\vec{\lambda}_0) + n$ and the likelihood is given by

$$L'(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\left\langle s - h(\vec{\lambda})\middle| s - h(\vec{\lambda})\right\rangle\right)\,, \tag{2}$$

Where $\langle\cdot|\cdot\rangle$ denotes the usual noise-weighted inner product

$$\left\langle a|b\right\rangle = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}\mathrm{d}f\,. \tag{3}$$

In Eq. (3), $S_n(f)$ is the (two-sided) noise power spectral density in the detector.

In general we do not have access to the true waveform model $h(\vec{\lambda})$, at least not at a reasonable computational cost. Highly, but not totally, accurate NR waveforms have recently started to become available [12], and slightly less accurate (but computationaly cheaper) extended analytic models such as (S)EOBNR [13] are also available. However, these are too computationally expensive to use in routine parameter estimation studies, which typically require many thousands of likelihood evaluations. Instead, we must make use of cheaper but less accurate waveforms, such as post-Newtonian (PN) [14]), or numerical "kludge" models [15]. Denoting the approximate waveform model by $H(\vec{\lambda})$, the *approximate likelihood* obtained when using this model is given by

$$L(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\left\langle s - H(\vec{\lambda})\middle| s - H(\vec{\lambda})\right\rangle\right)\,. \tag{4}$$

In general, posterior distributions obtained from this likelihood will not agree with posterior distributions obtained from the exact likelihood in Eq. (2). Denote by $\vec{\lambda}_{\rm exact}$ the best fit parameters obtained from Eq. (2) and $\vec{\lambda}_{\rm approx}$ the best fit parameters obtained from Eq. (4). If both the waveform difference and the parameter shift $\Delta\vec{\lambda} \equiv \vec{\lambda}_{\rm approx} - \vec{\lambda}_{\rm exact}$ are small quantities, $\mathcal{O}(\epsilon)$, then an approximate expression for the shift in the parameters can be found by expanding in ϵ. The shift in best-fit parameters to linear order in ϵ was obtained in [8] as $\Delta\vec{\lambda} \equiv \Delta\vec{\lambda}_1$ where

$$\Delta\lambda^a_1 = -\left(\Sigma^{-1}\right)^{ab}\left\langle \delta h(\vec{\lambda}_0)|\partial_b H(\vec{\lambda}_0)\right\rangle\,, \tag{5}$$

$\Sigma_{ab} = \left\langle \partial_a H(\vec{\lambda})|\partial_b H(\vec{\lambda})\right\rangle$, and $\partial_a = \partial/\partial\lambda^a|_{\vec{\lambda}=\vec{\lambda}_0}$. For completeness we include a derivation of this result, and an extension of it to quadratic order, in Appendix A).

From Eqs. (5) and (A6) it can be seen that the systematic shift in parameters caused by using the approximate likelihood is independent of the signal-to-noise ratio (SNR). This fact was observed in [8], and since the statistical errors that arise from detector noise decrease with increasing SNR this means that the systematic shift is most important for the loudest sources.

When using the approximate likelihood in Eq. (4) to characterise a single source one would usually use the condition that the systematic error due to the model uncertainty is less than the random error arising from noise to determine if the model is "good enough". This condition ensures that the true parameters will be consistent with the posterior — the amount by which the systematic error shifts the peak of the posterior is less than the typical posterior width. However, whilst this condition ensures that the true parameters will always be consistent with the posterior, on average they will be further from the centre of the posterior and hence lie at a lower significance than they should. This starts to become important when observing a population of sources (as we hope will be the case for Advanced LIGO). Even small systematic shifts may lead one to make incorrect inferences about the properties of the population. This can be understood by imagining that we observe a NS-NS binary with identical astrophysical parameters n independent times with Advanced LIGO. The error in the combined estimate for the mean mass of the population is the error in each measurement divided by $\sqrt{n}$. Therefore even if the systematic model error is insignificant for making inferences regarding a single binary it becomes increasingly significant for inferences regarding populations as new sources are added. The importance of the model errors for LIGO observations of NS-NS binaries was considered by [16]. Model error effects could also be seen in the parameter estimation analysis of the "big-dog" blind injection. In that case, the recovered masses for the compact binary injection were significantly biased (in part) by the fact that different signal models were used for the injection and parameter estimation [17]. This indicates the importance of considering how to incorporate model uncertainties in parameter estimation before the advanced detector era begins. A detailed investigation of parameter estimation on various injections into data from the LIGO/Virgo interferometers and employing a range of different models for the analysis was carried out in [18]. These results clearly show how the analysis of the same data using two different models can give mutually inconsistent results.

The recently proposed marginalised likelihood ([11]) attempted to account for the systematic error in the posterior, and hence remove the bias. The approximate likelihood is constructed by including information from a small training set of accurate waveforms computed offline;

$$\mathcal{D} = \left\{ (\vec{\lambda}_i, \delta h(\vec{\lambda}_i)) | i = 1, 2, \ldots, n \right\} , \tag{6}$$

in which $\delta h(\vec{\lambda}) \equiv H(\vec{\lambda}) - h(\vec{\lambda})$ denotes the difference between the approximate waveform and the true waveform. GPR assumes that the waveform differences in the training set are a realisation of a Gaussian process with covariance function $k(\vec{\lambda}, \vec{\lambda}')$ over the parameter space $\vec{\lambda}$. Different covariance functions may be considered and the *evidence* for the Gaussian process can be maximised with respect to variations in the parameters of the covariance function: this process of optimising the covariance function is called "training", and it enables the Gaussian process to "learn" the properties of the waveform differences in $\mathcal{D}$. The Gaussian process, once trained, may then be used to interpolate the waveform difference across parameter space. As we are not interested in the actual waveform difference, but rather in its effect on the posterior, the GPR interpolation is used as a prior to analytically marginalise over the unknown waveform difference. The resulting expression for the marginalised likelihood is [11]

$$\mathcal{L}(\vec{\lambda}) \propto \frac{\exp\left(-\frac{1}{2} \frac{\left\langle s - H(\vec{\lambda}) + \mu(\vec{\lambda})) \middle| s - H(\vec{\lambda}) + \mu(\vec{\lambda})) \right\rangle}{1 + \sigma^2(\vec{\lambda})} \right)}{\sqrt{1 + \sigma^2(\vec{\lambda})}} , \tag{7}$$

where the GPR quantity $\mu(\vec{\lambda})$ is the mean waveform difference and $\sigma^2(\vec{\lambda})$ is the error in this GPR estimate;

$$\mu(\vec{\lambda}) = k(\vec{\lambda}_i, \vec{\lambda}) \text{ inv} \left(k(\vec{\lambda}_i, \vec{\lambda}_j) \right) \ \delta h(\vec{\lambda}_j) , \tag{8}$$

$$\sigma^2(\vec{\lambda}) = k(\vec{\lambda}, \vec{\lambda}) - k(\vec{\lambda}_i, \vec{\lambda}) \text{inv} \left(k(\vec{\lambda}_i, \vec{\lambda}_j) \right) k(\vec{\lambda}_j, \vec{\lambda}). \tag{9}$$

For more details on the technique of Gaussian process regression see (for example) [19, 20] and for more details of the marginalised likelihood see [11].

III. ANALYTIC CALCULATION OF THE P-P PLOT

In the limit of high SNR the posterior probability distribution obtained in the analysis of data from a detector will be strongly peaked in the vicinity of the true parameters. Within the vicinity of this peak it is reasonable to expand both the exact and approximate signal models in the usual linear signal approximation (LSA), i.e.

$$\begin{aligned} h(\vec{\lambda}) &= h(\vec{\lambda}_0) + \Delta\vec{\lambda}^a \partial_a h(\vec{\lambda}_0) , \\ H(\vec{\lambda}) &= H(\vec{\lambda}_0) + \Delta\vec{\lambda}^a \partial_a H(\vec{\lambda}_0) . \end{aligned} \tag{10}$$

where $\vec{\lambda}_0$ denotes the parameter values of the true signal, $\vec{\lambda}$ denotes the parameter values at which we want to evaluate the signal or likelihood and $\Delta\vec{\lambda} = \vec{\lambda} - \vec{\lambda}_0$. This LSA is the usual approximation made in the derivation of the Fisher Matrix and the approximation used in the derivation of Eqs. (5) and (A6).

We are interested in predicting the "sag" that would be expected in a P-P plot. If we use an approximate waveform model to compute the posterior, then we would expect some bias in the recovered parameters and a sag in the P-P plot - on average the true parameters would be further away from the peak of the posterior than we would expect, and so fewer injections would be recovered at a given significance level.

A. The exact likelihood

The exact likelihood, by definition, will give a diagonal unbiased P-P plot. However we will re-derive this obvious result to shed light on the calculations that follow.

The *Exact Likelihood* is given by Eq. (2). The measured data is assumed to consist of a signal with true parameters $\vec{\lambda}_0$ and additive Gaussian noise; $s = h(\vec{\lambda}_0) + n$. In the limit of high SNR, the difference between two nearby signals in parameter space may be expanded using the LSA,

$$\begin{aligned} L'(\vec{\lambda}) &\propto \exp\left(-\frac{1}{2}\left\langle n - \Delta\lambda^a \partial_a h \middle| n - \Delta\lambda^a \partial_a h\right\rangle\right), \qquad (11)\\ &= \exp\left(-\frac{1}{2}\left[\langle n|n\rangle - 2\Delta\lambda^a \left\langle n\middle|\partial_a h\right\rangle + \Delta\lambda^a \Delta\lambda^b S_{ab}\right]\right), \end{aligned}$$

where the exact Fisher matrix is $S_{ab} = \langle \partial_a h | \partial_b h \rangle$. Since the Fisher matrix is symmetric by construction, we may adopt new coordinates in parameter space $\tilde{\Delta\lambda}^a = Q^a_b \Delta\lambda^b$ such that the Fisher matrix in these coordinates becomes diagonal, $S_{ab} = Q^p_a Q^q_b \delta_{pq}$. This amounts to rescaling the coordinate axes such that the iso-probability contour, which originally was an n-ellipsoid, becomes an n-sphere. Derivatives with respect to the new coordinates will be denoted with a tilde, $\partial_a h = Q^b_a \tilde{\partial}_b h$. In these new coordinates the likelihood separates to become

$$L'(\vec{\lambda}) \propto \prod_x \exp\left(-\frac{1}{2}\left(\tilde{\Delta\lambda}^x - \left\langle n\middle|\tilde{\partial}_x h\right\rangle\right)^2\right). \qquad (12)$$

In order to exploit the spherical symmetry about the peak in the rescaled parameters we adopt (n-dimensional) spherical coordinates centred on the peak; the radial coordinate given by $r^2 = \sum_x (\tilde{\Delta\lambda}^x - < n|\tilde{\partial}_x h >)^2$. The significance of the true parameters is given by the volume of the posterior that is "closer to the peak", i.e., that has higher posterior weight than the true parameters,

$$\begin{aligned} \text{sig} &= \frac{\int_0^R \mathrm{d}r\; r^{N-1}\exp(-r^2/2)}{\int_0^\infty \mathrm{d}r\; r^{N-1}\exp(-r^2/2)}\\ &= 1 - \frac{\Gamma\left(\frac{N}{2}, \frac{R^2}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} = 1 - \bar{\Gamma}\left(\frac{N}{2}, \frac{R^2}{2}\right), \qquad (13) \end{aligned}$$

where $\Gamma(x, y)$ is the incomplete Gamma function,

$$\Gamma(x, y) = \int_y^\infty t^{x-1}\mathrm{e}^{-t}\mathrm{d}t\,, \qquad (14)$$

$\Gamma(x) = \Gamma(x, 0)$ is the complete Gamma function, and $\bar{\Gamma}(x, y)$ is the regularised incomplete gamma function defined via the last equality in Eq. (13). In Eq. (13) the assumption has been made that the prior distribution on the parameters may be approximated as a constant across the width of the peak; this is reasonable in the high SNR limit when the posterior is narrow. The quantity R^2 is given by

$$R^2 = \sum_x \left\langle n\middle|\tilde{\partial}_x h\right\rangle^2 = \left(S^{-1}\right)^{ab} \left\langle n\middle|\partial_a h\right\rangle \left\langle n\middle|\partial_b h\right\rangle, \qquad (15)$$

and is distributed as a χ^2 random variable with $N = \dim(\vec{\lambda})$ degrees of freedom. The inverse regularised incomplete gamma function is defined via $y = \bar{\Gamma}(x, \bar{\Gamma}^{-1}(x, y))$. The quantity on the ordinate axis of a a standard P-P plot is the probability that the true parameters lie within a given significance, $P(\text{sig} < X)$. From Eq. (15) it may be seen that this can be rewritten as a cumulative probability of the random variable R^2;

$$P(\text{sig} < X) = 1 - P\left(R^2 < 2\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1 - X\right)\right). \qquad (16)$$

The cumulative distribution function of the χ^2 distribution is the regularised Gamma function, $P(R^2 < y) = \bar{\Gamma}(N/2, y/2)$. Using this to evaluate Eq. 16 gives the expected, unbiased diagonal form of the P-P plot for the exact likelihood;

$$P(\text{sig} < X) = 1 - (1 - X) = X\,. \qquad (17)$$

This diagonal P-P plot is shown in the dotted black curve in the left-hand panel of Fig. 1. The fact that the PP plot for the exact likelihood is always diagonal follows from the definition of the likelihood, and this remains true even if the LSA fails. The derivation just presented assumes the LSA in order to make it resemble as closely as possible the upcoming derivation for the approximate likelihood.

B. The approximate likelihood

We now move on to the more interesting case when we have biased parameter estimation from using the approximate likelihood. As mentioned in the introduction we expect to obtain a P-P plot that is "sagging" below the diagonal indicating the bias. We first treat the simple case where the waveform model depends on just a single parameter, $\vec{\lambda} = \theta$, where the expression for the P-P plot is given in terms of the inverse error function, $\text{erf}^{-1}(x)$. A treatment will then be given for the general N dimensional case in which the expression for the P-P plot is given in terms of the MarcumQ function, $Q_N(x, y)$, along with a illustration of how this reduces to the 1D result.

The *Approximate Likelihood* is given by Eq. (4). We assume the approximate model is "nearly" correct and use the LSA to expand signals that are nearby in parameter space. As before, denoting the waveform difference

by $\delta h(\vec{\lambda}) = H(\vec{\lambda}) - h(\vec{\lambda})$, we have

$$\begin{aligned} L(\vec{\lambda}) &\propto \exp\left(-\frac{1}{2}\left\langle n - \delta h(\vec{\lambda}_0) - \Delta\lambda^a \partial_a H \middle| \ldots \right\rangle\right) \\ &= \exp\left(-\frac{1}{2}\left[\left\langle n - \delta h(\vec{\lambda}_0)\middle| \ldots \right\rangle\right.\right. \\ &\quad \left.\left. -2\Delta\lambda^a \left\langle n - \delta h(\vec{\lambda}_0)|\partial_a H\right\rangle + \Delta\lambda^a\Delta\lambda^b \Sigma_{ab}\right]\right), \end{aligned} \tag{18}$$

where the ellipsis in the right hand entry in the inner product denotes a repeat of the left hand entry and the approximate Fisher matrix is $\Sigma_{ab} = \langle \partial_a H | \partial_b H \rangle$. As before coordinates which diagonalise the Fisher matrix Σ_{ab} may be adopted, which give the following separated expression for the approximate likelihood,

$$L(\vec{\lambda}) \propto \prod_x \exp\left(-\frac{1}{2}\left(\Delta\tilde{\lambda}^x - \left\langle n - \delta h(\vec{\lambda}_0)\middle|\tilde{\partial}_x H\right\rangle\right)^2\right). \tag{19}$$

1. Example for one dimensional parameter space

If the waveform depends on only one unknown parameter, $\vec{\lambda} = \lambda$, Eq. (19) becomes

$$L(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\Delta\theta - \mu\right)^2\right], \tag{20}$$

where

$$\frac{1}{\sigma^2} = \left\langle \frac{\mathrm{d}H}{\mathrm{d}\lambda}\Big|_{\lambda=\lambda_0} \middle| \frac{\mathrm{d}H}{\mathrm{d}\lambda}\Big|_{\lambda=\lambda_0} \right\rangle, \tag{21}$$

$$\mu = \sigma^2 \left\langle n - \delta h(\lambda_0)\middle| \frac{\mathrm{d}H}{\mathrm{d}\lambda}\Big|_{\lambda=\lambda_0} \right\rangle, \tag{22}$$

and we have included the correct normalisation of the posterior. The true parameter value is at $\Delta\theta = 0$ and the points with larger posterior weight than the true parameters lie in the range $0 < \Delta\theta < 2\mu$ when $\mu > 0$ or in the range $2\mu < \Delta\theta < 0$ when $\mu < 0$. The significance at which the true parameters lie is therefore

$$\int_0^{2\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\Delta\theta - \mu\right)^2\right] \mathrm{d}\Delta\theta = \mathrm{erf}\left(\frac{|\mu|}{\sqrt{2}\sigma}\right), \tag{23}$$

where

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-t^2} \mathrm{d}t \tag{24}$$

is the usual error function. The quantity μ defined above depends on the particular realisation of the noise. We want to know the fraction of times, over many realisations of the noise, that the true parameters will lie within a certain significance contour. This is just

$$P(\mathrm{sig} < X) = P\left(\frac{|\mu|}{\sqrt{2}\sigma} < \mathrm{erf}^{-1}(X)\right). \tag{25}$$

The quantity $\mu/(\sqrt{2}\sigma)$ is distributed as a Gaussian with mean $\tilde{\mu} = \sigma \left\langle \Delta h(\lambda_0) | \mathrm{d}H/\mathrm{d}\lambda \right\rangle / \sqrt{2}$ and variance 1/2 and so

$$\begin{aligned} P(\mathrm{sig} < X) = \frac{1}{2}&\mathrm{erf}\left(\mathrm{erf}^{-1}(X) - \tilde{\mu}\right) \\ &+\frac{1}{2}\mathrm{erf}\left(\mathrm{erf}^{-1}(X) + \tilde{\mu}\right). \end{aligned} \tag{26}$$

In the special case where the approximate waveform model and the exact waveform model are the same, we have $\tilde{\mu} = 0$ and recover the expected unbiased result from Sec. III A;

$$P(\mathrm{sig} < X) = X. \tag{27}$$

This derivation assumed that $\tilde{\mu}$ was constant, but in practice this will vary from signal to signal. If we denote the probability distribution function for $\tilde{\mu}$ over the astrophysical population by $f(\tilde{\mu})$, the generalisation of Eq. (26) can be seen straightforwardly to be

$$\begin{aligned} P(\mathrm{sig} < X) = \int \Bigg[&\frac{1}{2}\mathrm{erf}\left(\mathrm{erf}^{-1}(X) - \tilde{\mu}\right) \\ &+\frac{1}{2}\mathrm{erf}\left(\mathrm{erf}^{-1}(X) + \tilde{\mu}\right)\Bigg] f(\tilde{\mu}) \mathrm{d}\tilde{\mu}. \end{aligned} \tag{28}$$

2. Parameter space of arbitrary dimension

We will now generalise the expression for the P-P plot sag in a one-dimensional parameter space, given in Eq. (26), to arbitrary numbers of parameters. Identical manipulations to those performed on the exact likelihood yields the same expression for the significance obtained in Eq. (13),

$$\mathrm{sig} = 1 - \bar{\Gamma}\left(\frac{N}{2}, \frac{R^2}{2}\right), \tag{29}$$

except this time R^2 is a random variable given by

$$\begin{aligned} R^2 &= \sum_x \left\langle n - \delta h(\vec{\lambda}_0)\middle|\tilde{\partial}_x H\right\rangle^2 \\ &= \left(\Sigma^{-1}\right)^{ab} \left\langle n - \delta h(\vec{\lambda}_0)\middle|\partial_a H\right\rangle \left\langle n - \delta h(\vec{\lambda}_0)\middle|\partial_b H\right\rangle. \end{aligned} \tag{30}$$

If $\delta h(\vec{\lambda})$ is constant across parameter space, R^2 is now a non-central χ^2 random variable with N degrees of freedom and non-centrality parameter

$$\Lambda = \left(\Sigma^{-1}\right)^{ab} \left\langle \delta h(\vec{\lambda}_0)\middle|\partial_a H\right\rangle \left\langle \delta h(\vec{\lambda}_0)\middle|\partial_b H\right\rangle. \tag{31}$$

As before the expression for the P-P plot is given in terms of the CDF of the distribution of the random variable R^2. The CDF of the non-central χ^2 distribution is the Marcum-Q function, $P(R^2 < y) = Q_{N/2}(\sqrt{\Lambda}, \sqrt{y})$,

$$P(\mathrm{sig} < X) = 1 - P\left(R^2 < 2\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1 - X\right)\right) \tag{32}$$

$$= 1 - Q_{\frac{N}{2}}\left(\sqrt{\Lambda}, \sqrt{2\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1 - X\right)}\right) \tag{33}$$

This is an analytic approximation to the P-P plot in the LSA and in the case of a constant waveform difference over parameter space; this function is plotted as a dotted black line in Fig. 1. In this case the P-P plot always sags below the diagonal indicating biased parameter recovery.

If $\delta h(\vec{\lambda})$ is not constant over parameter space, the generalisation of this result takes the same form as Eq. (28), but with the term in square brackets replaced by Eq. (33) and with $f(\tilde{\mu})$ replaced by the corresponding probability distribution function for Λ. For example, in the case that $\delta h(\vec{\lambda})$ is distributed at different times and at different points in parameter space as an uncorrelated, zero-mean Gaussian with variance in each component of ϵ^2 (i.e. $\delta h(\vec{\lambda}_0) \sim \mathcal{N}(0, \epsilon^2)$) then the quantities $\langle \delta h(\vec{\lambda}_0) | \partial_a H \rangle$ are distributed as $N(0, \Sigma)$ and we see that Λ is distributed as ϵ^2 times a χ^2 distribution with N degrees of freedom with probability distribution function

$$f(\Lambda) = \frac{1}{2^{\frac{N}{2}}\Gamma(N/2)\epsilon^N}\Lambda^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\Lambda}{2\epsilon^2}}. \tag{34}$$

Writing $x_u^2 = 2\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1-X\right)$ we must evaluate

$$\begin{aligned}
&P(\mathrm{sig} < X) \\
&= \int_0^\infty \left[\frac{\Lambda^{\frac{N}{2}-1}\mathrm{e}^{-\frac{\Lambda}{2\epsilon^2}}}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)\epsilon^N} \right. \\
&\qquad \left. \int_0^{x_u} x\left(\frac{x}{\sqrt{\Lambda}}\right)^{\frac{N}{2}-1} \mathrm{e}^{-\frac{1}{2}(x^2+\Lambda)} I_{\frac{N}{2}-1}(\sqrt{\Lambda}x)\mathrm{d}x \right] \mathrm{d}\Lambda \\
&= \frac{1}{(2\epsilon)^{\frac{N}{2}-1}\Gamma\left(\frac{N}{2}\right)} \int_0^{x_u} \left[x^{\frac{N}{2}}\mathrm{e}^{-\frac{x^2}{2}} \right. \\
&\qquad \left. \int_0^\infty y^{\frac{N}{2}}\mathrm{e}^{-\frac{1}{2}(1+\epsilon^2)y^2} I_{\frac{N}{2}-1}(\epsilon x y)\mathrm{d}y \right] \mathrm{d}x \\
&= \frac{1}{(2\epsilon)^{\frac{N}{2}-1}\Gamma\left(\frac{N}{2}\right)(1+\epsilon^2)^{\frac{1}{2}+\frac{N}{4}}} \int_0^{x_u} \left[x^{\frac{N}{2}}\mathrm{e}^{-\frac{x^2}{2}} \right. \\
&\qquad \left. \int_0^\infty \tilde{y}^{\frac{N}{2}}\mathrm{e}^{-\frac{1}{2}\tilde{y}^2} I_{\frac{N}{2}-1}\left(\frac{\epsilon}{\sqrt{1+\epsilon^2}}x\tilde{y}\right) \mathrm{d}\tilde{y} \right] \mathrm{d}x \\
&= \frac{1}{2^{\frac{N}{2}-1}\Gamma\left(\frac{N}{2}\right)(1+\epsilon^2)^{\frac{N}{2}}} \int_0^{x_u} x^{N-1}\mathrm{e}^{-\frac{x^2}{2(1+\epsilon^2)}}\mathrm{d}x \\
&= \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} \int_0^{\frac{x_u^2}{1+\epsilon^2}} u^{\frac{N}{2}-1}\mathrm{e}^{-\frac{u}{2}}\mathrm{d}u \\
&= 1 - \bar{\Gamma}\left(\frac{N}{2}, \frac{\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1-X\right)}{1+\epsilon^2}\right),
\end{aligned} \tag{35}$$

where the second line follows by a change of variable and a change in the order of integration, the fourth line follows from the fact that $Q_m(a, 0) = 1$ and the final lines follow from another change of variable. We have also made use of the integral expression for the Marcum-Q function given below. This result can also be obtained directly by noticing that the random variable R^2, which depends on both n and $\delta h(\vec{\lambda})$, is distributed as $1+\epsilon^2$ times a χ^2 random variable with N degrees of freedom. The analytic expression for the P-P plot is therefore very similar to the case of the exact likelihood, but with the argument of the regularised Gamma function scaled appropriately to give the same result as in the final line of Eq.(35). This P-P plot also exhibits a sag below the diagonal; see the orange dotted curve in Fig. 1.

The N dimensional result in Eq. (33) can be shown to reduce to the 1 dimensional result in Eq. (26) using the standard properties of the Marcum-Q function. The Marcum-Q function is defined by the integral

$$\begin{aligned}
Q_m(a, b) &= \int_b^\infty x\left(\frac{x}{a}\right)^{m-1} \exp\left[-\frac{1}{2}(x^2+a^2)\right] I_{m-1}(ax)\mathrm{d}x \\
&= \exp\left[-\frac{1}{2}(a^2+b^2)\right] \sum_{k=1-m}^{\infty} \left(\frac{a}{b}\right)^k I_k(ab),
\end{aligned} \tag{36}$$

in which $I_n(x)$ is the modified Bessel function of the first kind. For $N = 1$, the Marcum-Q function, Eq. (36), can also be simplified

$$\begin{aligned}
Q_{\frac{1}{2}}(a, b) &= \sqrt{a}\int_b^\infty \sqrt{x}\exp\left[-\frac{(x^2+a^2)}{2}\right] I_{-1/2}(ax)\mathrm{d}x \\
&= \sqrt{\frac{1}{2\pi}}\int_b^\infty \left(\exp\left[-\frac{(x+a)^2}{2}\right] + \exp\left[-\frac{(x-a)^2}{2}\right]\right)\mathrm{d}x \\
&= 1 - \frac{1}{2}\left(\mathrm{erf}\left(\frac{b-a}{\sqrt{2}}\right) + \mathrm{erf}\left(\frac{b+a}{\sqrt{2}}\right)\right),
\end{aligned} \tag{37}$$

which follows from $I_{-\frac{1}{2}}(x) = \sqrt{2/\pi}\cosh(x)/\sqrt{x}$. When $N = 1$ the regularised Gamma function becomes

$$\begin{aligned}
\frac{\Gamma\left(1/2, R^2/2\right)}{\Gamma\left(1/2\right)} &= \frac{\int_{R^2/2}^\infty \mathrm{e}^{-t}/\sqrt{t}\ \mathrm{d}t}{\int_0^\infty \mathrm{e}^{-t}/\sqrt{t}\ \mathrm{d}t} \\
&= \frac{\int_{R/\sqrt{2}}^\infty \mathrm{e}^{-u^2}\ \mathrm{d}u}{\int_0^\infty \mathrm{e}^{-u^2}\ \mathrm{d}u} \\
&= 1 - \mathrm{erf}(R/\sqrt{2}).
\end{aligned} \tag{38}$$

Eq. (32) therefore becomes

$$\begin{aligned}
P(\mathrm{sig} < X) = \frac{1}{2}\left(\mathrm{erf}\left(\mathrm{erf}^{-1}(\mathrm{X}) - \sqrt{\frac{\Lambda}{2}}\right)\right. \\
\left. + \mathrm{erf}\left(\mathrm{erf}^{-1}(\mathrm{X}) + \sqrt{\frac{\Lambda}{2}}\right)\right),
\end{aligned} \tag{39}$$

as we can identify $\tilde{\mu} = \Lambda/2$, we recover Eq. (26) as expected.

C. The marginalised likelihood

The *Marginalised Likelihood* is given by Eq. (7), as before this may be expanded in the LSA. In the high SNR limit the posterior is narrow compared to the length scale over which the waveform changes. The waveform difference changes over the same length scale as the waveform. The quantity $\sigma^2(\vec{\lambda})$ also changes over this length scale,

as it is "learnt" by the GP in the procedure of maximising the evidence. Therefore in the high SNR limit $\sigma^2(\vec{\lambda})$ may be approximated as a constant. As before coordinates which diagonalise the Fisher matrix may be adopted, which give the following separated expression for the approximate likelihood,

$$\mathcal{L}(\vec{\lambda}) \propto \prod_x \exp\left(-\frac{1}{2}\frac{\left(\tilde{\Delta\lambda}^x - \left\langle n + \mu(\vec{\lambda}_0) - \delta h(\vec{\lambda}_0)\middle|\tilde{\partial}_x(H-\mu)\right\rangle\right)^2}{1+\sigma^2}\right) . \tag{40}$$

The waveform difference is assumed to be a small quantity, therefore in Eq. (40) the derivative $\tilde{\partial}_x(H-\mu)$ may be replaced by $\tilde{\partial}_x(H)$, as the difference is the product of small quantities. Identical manipulations to those performed on the exact and approximate likelihoods give the same expression for the significance obtained in Eqs. (13) and (29),

$$\text{sig} = 1 - \bar{\Gamma}\left(\frac{N}{2}, \frac{R^2}{2}\right) , \tag{41}$$

except this time the random variable R^2 is given by

$$R^2 = \frac{1}{1+\sigma^2}\sum_x \left\langle n + \mu(\vec{\lambda}_0) - \delta h(\vec{\lambda}_0)\middle|\tilde{\partial}_x H\right\rangle^2 , \tag{42}$$

$$= \frac{1}{1+\sigma^2}\left(\Sigma^{-1}\right)^{ab}\left\langle n + \mu(\vec{\lambda}_0) - \delta h(\vec{\lambda}_0)\middle|\partial_a H\right\rangle\left\langle n + \mu(\vec{\lambda}_0) - \delta h(\vec{\lambda}_0)\middle|\partial_b H\right\rangle . \tag{43}$$

The GPR technique assumes that the $\delta h(\vec{\lambda})$ are distributed as a Gaussian process across parameter space, with zero mean and a covariance estimated from a training set and any prior knowledge. If this assumption is in fact true, and the covariance has been correctly estimated, then the quantity $\mu(\vec{\lambda}_0) - \delta h(\vec{\lambda}_0)$ is distributed as a zero mean Gaussian with variance σ^2. In this case (perhaps unsurprisingly) the marginalised likelihood completely fixes the sag. The new R^2 random variable is distributed as a χ^2 random variable with N degrees of freedom and using the regularised Gamma function as the CDF of this distribution we recover the diagonal P-P plot;

$$P(\text{sig} < X) = 1 - P\left(R^2 < 2\bar{\Gamma}^{-1}\left(\frac{N}{2}, 1-X\right)\right) , \tag{44}$$

$$= X . \tag{45}$$

This case is shown, both analytically and numerically, in orange in Fig. 1.

More interesting is the behaviour in the realistic case when $\delta h(\vec{\lambda})$ is not distributed exactly as the GPR has predicted. This case is more complicated because the different components that make up the R^2 random variable are no longer independent random variables and a simple expression for the distribution of $\delta h(\vec{\lambda})$ cannot be found. In particular, from Eq. (42) it can be seen that R^2 is the sum of the squares of a noise term, $< n|\tilde{\partial}_x H >$, a GPR term, $< \mu(\vec{\lambda}_0)|\tilde{\partial}_x H >$, and (minus) a physical term, $< \delta h(\vec{\lambda}_0)|\tilde{\partial}_x H >$. In particular the GPR and physical terms are now related because the expression for $\mu(\vec{\lambda}_0)$ in Eq. (8) is a linear combination of the realisations of $\delta h(\vec{\lambda})$ in the training set, $\mathcal{D}$. The sag will still be given by the analogue of Eq. (28), but this integral will not in general be analytically tractable. Instead, we will consider such cases numerically in Sec. IV.

As we have seen, in the particular case considered above where the waveform difference is distributed as assumed by the GPR, the marginalised likelihood completely removes the systematic bias present in the standard, approximate, likelihood. In addition, as we will see in Sec. IV, even in unfavourable situations the marginalised likelihood is often able to remove significant portions of the bias. We conclude this section with a discussion of why it is expected that the bias in parameter estimates obtained using the marginalised likelihood will usually be less than those obtained using the standard likelihood.

From Eqs. (30) and (42) it can be seen that the condition for the marginalised likelihood to yield *more* biased parameter estimates than the approximate likelihood, *for a particular event*, is $R^2_{\text{approx}} < R^2_{\text{GPR}}/(1+\sigma^2(\vec{\lambda}_0))$, where

$$R^2_{\text{approx}} = \sum_x \left\langle n - \delta h(\vec{\lambda}_0)\middle|\tilde{\partial}_x H\right\rangle^2$$

$$R^2_{\text{GPR}} = \sum_x \left\langle n - \delta h(\vec{\lambda}_0) + \mu(\vec{\lambda}_0)\middle|\tilde{\partial}_x H\right\rangle^2 . \tag{46}$$

These terms both involve a projection onto the space spanned by the derivatives, $\tilde{\partial}_x H$, at the point $\vec{\lambda}_0$. Since these "tilde" derivatives were constructed to be an orthonormal basis, the condition for the marginalised likelihood to give worse parameter estimates than the approximate likelihood can therefore be written as

$$\left|n - \delta h(\vec{\lambda}_0)\right|^2_{\mathcal{D}} < \frac{\left|n - \delta h(\vec{\lambda}_0) + \mu(\vec{\lambda}_0)\right|^2_{\mathcal{D}}}{1 + \sigma^2(\vec{\lambda}_0)} \qquad (47)$$

where the modulus is taken with respect to the function inner product in Eq. (3), projected into the space, $\mathcal{D}$, spanned by the derivatives. For this to be satisfied, it would be necessary not only for the interpolation to have the wrong sign when expressed in the basis $\tilde{\partial}_x H$ (i.e. $0 > \sum_x \langle\, h(\vec{\lambda}_0)|\tilde{\partial}_x H\,\rangle\,\langle\,\mu(\vec{\lambda}_0)|\tilde{\partial}_x H\,\rangle$), but also for it to be large enough in magnitude to overcome the GPR uncertainty $\sigma^2(\vec{\lambda}_0)$ in the denominator. Moreover, this is just for one particular realisation of the noise and true waveform parameters. We are really interested in the sag that arises when considering a population of events. In that case, we would need Eq. (47) to be true in some average sense and so the interpolation would have to have the wrong sign and be too large for the majority of choices of waveform parameters. Although this is technically possible, it is clear that any reasonable interpolation algorithm with decent coverage of the parameter space in the training set and a reasonable covariance function should violate the above bound on average and therefore yield better parameter estimates on average and show a smaller sag in the P-P plot than the approximate likelihood.

IV. NUMERICAL CALCULATION OF THE P-P PLOT

In all of the above calculations the expression for the P-P plot was written in terms of the CDF of the distribution of the R^2 random variable. This random variable is written in terms of a signal inner product of the model derivatives, it therefore depends both on the properties of the GW source and of the GW detector. By expressing our results in terms of R^2 we ensure that they remain valid for any detector and any source (assuming the LSA holds). In the cases considered above where analytic expression for the P-P plots could be found these can also be verified numerically by drawing n values of R^2 from the relevant distribution and numerically estimating the CDF. In cases where an analytic expression for the P-P plot can not be found the same procedure can be used to investigate the P-P plot numerically.

First consider the unbiased, diagonal P-P plot obtained for the exact likelihood. The analytic expression for this P-P plot is given in Eq. (17). A numerical validation of this result may be performed by drawing random realisations of the R^2 value in Eq. (15). It can be seen that R^2 is the sum of the squares of N standard Gaussian random variables $< n|\tilde{\partial}_x H >$; i.e. a χ^2 distribution with N degrees of freedom. We drew n realisations of R^2 from this distribution, numerically estimated the CDF and plotted the P-P plot using Eq. (16). The results for $n = 10^3$ and $N = 4$ are shown in the left panels of Fig. 1 (analytic results shown as a dotted line, numerical results as a solid line). Within the scale of fluctuations the numerical results agree well with the analytic results. The bottom left panel of the same figure shows the sag of the P-P plot below the diagonal, i.e. $\mathrm{sig} - P(x < \mathrm{sig})$. The values $n = 10^3$ and $N = 4$ will also be used for all subsequent numerical calculations in this section.

P-P plots for the approximate likelihood are shown in the centre panels of Fig. 1 for a variety of different distributions of the waveform difference projected into the model derivatives; $< \delta h(\lambda_0)|\tilde{\partial}_x H >$. In the case of a constant distribution, or a zero-mean Gaussian distribution the analytic expressions in Eqs. (33) and (35) respectively are shown as dotted lines. For the numerical calculations the procedure followed was first to specify the distribution for the $< \delta h(\lambda_0)|\tilde{\partial}_x H >$ random variables (for example the black curves show results when this is a constant). The quantity R^2 was then calculated using Eq. (30) by drawing a random value from this distribution and a random value for $< n|\tilde{\partial}_x H >$ from a standard Gaussian distribution. The R^2 variable was calculated n times, the CDF of this variable estimated, and the P-P plot calculated from Eq. (32). Different colours in Fig. 1 indicate different distributions for $< \delta h(\lambda_0)|\tilde{\partial}_x H >$, the specification of these distributions are given in the figure caption.

	Approximate	Marginalised
Constant	0.158	-0.044
Gaussian	0.237	0.000
non-central Gaussian	0.385	0.263
Skew non-central Gaussian	0.426	0.079
Poisson	0.317	0.235
Gamma	0.293	-0.001
Correlated	0.441	0.308

TABLE I. Table of the total integrated biases for the curves shown in Fig. 1. The integrated bias is defined as the total area in the sag, i.e. $\int_0^1 \mathrm{d(sig)}\ (\mathrm{sig} - P(x < \mathrm{sig}))$.

P-P plots for the marginalised likelihood are shown in the right panels of Fig. 1 for a variety of different distributions of $< \delta h(\lambda_0)|\tilde{\partial}_x H >$. In the case of a zero-mean Gaussian the analytic expressions in Eq. (45) is shown as a dotted line. For the numerical calculations it is necessary to construct a training set for the GPR interpolation. Instead of using GPR to interpolate the waveform differences, $\delta h(\vec{\lambda})$, it is simpler for our present purpose to instead interpolate the projections of the waveform differences onto the waveform derivatives, i.e., $< \delta h(\vec{\lambda})|\tilde{\partial}_x H >$, as these are what appear in Eq. (42), . The training set was taken to consist

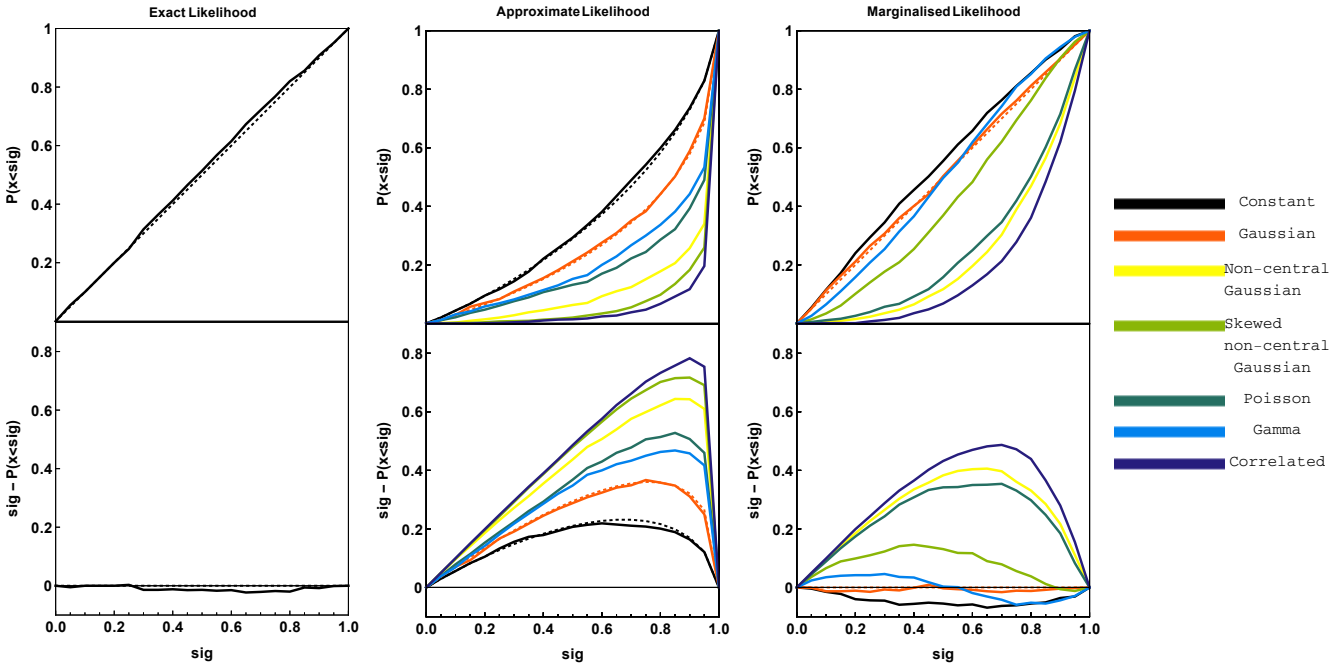

FIG. 1. P-P plots for parameter estimation using the three likelihoods $L'(\vec{\lambda})$, $L(\vec{\lambda})$, and $\mathcal{L}(\vec{\lambda})$ shown in the three columns respectively. In each column the top panel shows a P-P plot whilst the bottom panel shows the "sag"; i.e. the difference between the ideal diagonal line and the actual P-P plot. In each panel curves drawn as dotted lines correspond to analytic results whilst solid curves are numerical results. The left-hand column shows ideal, unbiased parameter recovery for the exact likelihood. In the centre and right-hand columns different colour curves correspond to different distributions of $< \delta h(\lambda_0)|\tilde{\partial}_x H >$. The curves in black are for a constant distribution giving a non-centrality $\Lambda = 2$ (Λ defined in Eq. (31)). The curves in orange are for a zero mean Gaussian distribution with variance 1. The curves in yellow are for a non-central Gaussian distribution with mean 4/3 and variance 1. The curves in light-green are for a non-central, skewed normal distribution[a] with location parameter 1, scale parameter 1, and skew parameter 1. The curves in dark-green are for a Poisson distribution with mean and variance 1. The curves in blue are for a Gamma distribution with shape parameter 1 and scale parameter 1. And finally, the curves in purple are for a correlated random walk distribution with mean Gaussian step size 1. In all cases the number of parameter dimensions is $N = 4$, and the number of points used for the numerical simulations was $n = 10^3$. The left-hand panel clearly shows the exact likelihood does not suffer from any bias, as expected. The centre panel shows that in all cases the approximate likelihood suffers from a bias. The right-hand column shows that in all cases the marginalised likelihood reduces the bias relative to the approximate likelihood. In the ideal case (shown in orange) of a zero mean Gaussian distribution for $< \delta h(\lambda_0)|\tilde{\partial}_x H >$ the bias is completely removed.

[a] The PDF of a skew Gaussian distribution with location parameter μ, scale parameter σ and skew parameter α is given by $\left[1 + \mathrm{erf}(\alpha(x-\mu)/\sqrt{2}\sigma)\right] \exp(-(x-\mu)^2/2\sigma^2)$

of points at $\lambda = 1, 2, \ldots, 20$ and the actual experimental realisation at a value $\lambda_0 = 21$. For the majority of distributions (constant, Gaussian, non-central Gaussian, skew non-central Gaussian, Poisson, and Gamma distributions) shown in Fig. 1 the random variables in the training set were drawn independently and interpolated using an uncorrelated Gaussian process, i.e. $K_{ij} = \sigma_f \delta_{ij}$. The R^2 value was calculated from Eq. 42 (with $\mu = 0$ from Eq. (8), because of the assumption of an uncorrelated process), the CDF estimated and the P-P plot calculated from Eq. (44).

The assumption of an uncorrelated Gaussian process is a conservative assumption. In the absence of correlations the Gaussian process regression assumes a "worst-case" scenario and returns a mean waveform difference of zero (see Eq. (8)). If correlations were present then the GPR would return a non-zero estimate for μ and shift the position of the posterior peak into better agreement with the true value, thus improving the P-P plot. To investigate the effect of correlations the final numerical calculation (labelled as "correlated" in Fig. 1) was performed using a random walk distribution. The values of $< \delta h(\vec{\lambda})|\tilde{\partial}_x H >$ at the points $\lambda = 1, 2, \ldots, 21$ were taken to be a realisation of a random walk with Gaussian step width $a = 1/3$. The first 20 of these values were taken as the training set and used to extrapolate the final value. For the GPR interpolation a squared exponential covariance function $k(x, y) = \exp((-1/2)(x-y)^2)$ was used. The squared exponential covariance function is not able to accurately capture the covariance of the random walk distribution, so this again represents a conservative choice to examine how the marginalised likelihood performs in the presence

of un-modelled correlations. However, even in this unfavourable case the marginalised likelihood still significantly reduces the bias in the P-P plot.

The purpose of considering such a wide variety of different distributions for the waveform difference is to test whether the marginalised likelihood is robust against different types of errors in the waveform models, which are not correctly modelled by the Gaussian process. For example, the marginalised likelihood assumes the waveform difference is a zero mean Gaussian process across parameter space, therefore it is perhaps not surprising that it performs well in the case of a zero mean Gaussian distribution. However the list of distributions used here also test the robustness of the method against non-central distributions (e.g non-central Gaussian), skewed distributions (e.g. skewed Gaussian), one-sided and non-Gaussian distributions (e.g. Poisson or Gamma distributions), and the presence of un-modelled correlations in the waveform difference (the random walk distribution).

By comparing the curves of the same colour between the centre and right-hand panels of Fig. 1 it can be seen that in all cases the P-P plot for the marginalised likelihood exhibits less of a bias, i.e., less of a "sag", than the approximate likelihood. In the ideal case where the distribution of the waveform differences is precisely that assumed by the GPR, a diagonal, unbiased P-P plot is recovered; however the bias is also almost completely removed for several of the other distributions considered. In all cases a significant improvement in performance can be seen when using the marginalised likelihood in place of the standard approximate likelihood. These results are summarised in Table I, which lists the *total bias* (defined as the area between the sagging curve and the ideal diagonal) for all the curves shown in Fig. 1.

V. DISCUSSION

The P-P plot provides a way to quantify the bias that results when using inaccurate models to perform GW parameter estimation. For individual sources the systematic error in the parameters is independent of the SNR, whilst the random errors scale as 1/SNR, and hence the bias is most significant for the loudest sources. Even in cases where, for each individual source, the systematic error is small compared to the random error, the bias can still be significant when observing populations of sources, since the statistical error in a parameter estimated from combining a population of sources reduces as $1/\sqrt{N}$ as more sources are added, while the systematic errors remain fixed.

In this paper several analytic expressions have been obtained that predict the sag of the P-P plots that results from different distributions of the model error. These results have been derived within the linear signal approximation, and are valid to $\mathcal{O}(1/\mathrm{SNR})$. These analytic expressions for the P-P plots may be viewed in the same spirit as Fisher matrix estimates for the random errors, or Cutler and Vallisneri's [8] expression for the systematic error in a single measurement. This latter result has also here been generalised (in Appendix A) to include terms of $\mathcal{O}(1/\mathrm{SNR}^2)$.

It is now well established that model errors will present significant problems for a range of GW sources. The authors recently proposed a novel method for tackling this problem; using a modified likelihood constructed using Gaussian process regression on a training set of accurate waveforms. In this paper the performance of this marginalised likelihood was examined by comparing the P-P plots (obtained both analytically and numerically) with those obtained from the standard likelihood. In particular, it was found that in favourable conditions the marginalised likelihood was able to completely remove the parameter estimation bias. More importantly, it was found that the marginalised likelihood was robust against a wide range of un-modelled features in the distribution of waveform differences, and in all cases considered outperformed the standard likelihood. These results provide further illustration of the need to account for model uncertainties (using GPR or other techniques) when drawing inferences from near future GW observations.

[1] G. M. Harry (The LIGO Scientific Collabortation), Classical and Quantum Gravity **27**, 084006 (2010).

[2] F. Acernese et al. (The Virgo Collaboration), Virgo Technical Report VIR-0027A-09 (2009), URL `http://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0027A-09.pdf`.

[3] M. A. McLaughlin, Classical and Quantum Gravity **30**, 224008 (2013), 1310.0758.

[4] G. Hobbs, Classical and Quantum Gravity **30**, 224007 (2013), 1307.2629.

[5] M. Kramer and D. J. Champion, Classical and Quantum Gravity **30**, 224009 (2013).

[6] R. N. Manchester and IPTA, Classical and Quantum Gravity **30**, 224010 (2013), 1309.7392.

[7] P. A. Seoane, S. Aoudia, H. Audley, G. Auger, S. Babak, J. Baker, E. Barausse, S. Barke, M. Bassan, V. Beckmann, et al. (The eLISA Consotium) (2013), 1305.5720.

[8] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007), URL `http://link.aps.org/doi/10.1103/PhysRevD.76.104018`.

[9] T. Sidery, B. Aylott, N. Christensen, B. Farr, W. Farr, F. Feroz, J. Gair, K. Grover, P. Graff, C. Hanna, et al., Phys. Rev. D **89**, 084060 (2014), 1312.6013.

[10] M. Vallisneri and N. Yunes, Phys. Rev. D **87**, 102002 (2013), 1301.2627.

[11] C. J. Moore and J. R. Gair, Physical Review Letters **113**, 251101 (2014), 1412.3657.

[12] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005), URL `http://link.aps.org/doi/10.1103/PhysRevLett.95.121101`.

[13] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **84**, 124052 (2011), 1106.1021.
[14] L. Blanchet, Living Reviews in Relativity **17** (2014), URL `http://www.livingreviews.org/lrr-2014-2`.
[15] S. Babak, H. Fang, J. R. Gair, K. Glampedakis, and S. A. Hughes, Phys. Rev. D **75**, 024005 (2007), gr-qc/0607007.
[16] P. Canitrot, Phys. Rev. D **63**, 082005 (2001).
[17] J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, C. Affeldt, et al., Phys. Rev. D **85**, 082002 (2012), 1111.7314.
[18] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, T. Adams, et al., Phys. Rev. D **88**, 062001 (2013), 1304.1775.
[19] D. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).
[20] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006).

Appendix A: Systematic bias due to waveform errors

We assume that an approximate model $H(\vec{\lambda})$ is used to recover the parameters of a gravitational wave signal that is described by the true model $h(\vec{\lambda})$ with parameters $\vec{\lambda}_0$. The best fit parameters of the approximate model are $\vec{\lambda}_{\rm bf} = \vec{\lambda}_0 + \Delta\vec{\lambda}$. These parameters minimise the squared distance between the true and approximate model spaces,

$$\left\langle \delta h(\vec{\lambda}_0) + H(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_0) \middle| \delta h(\vec{\lambda}_0) + H(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_0) \right\rangle \tag{A1}$$

where $\delta h(\vec{\lambda}_0) = H(\vec{\lambda}_0) - h(\vec{\lambda}_0)$ (note the different sign convention from [8]). Differentiating with respect to each of the parameters in turn, we find that the best-fit parameters must satisfy the equations

$$\langle \delta h(\vec{\lambda}_0) + H(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_0) | \partial_a \left(H(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_0) \right) \rangle = 0\,. \tag{A2}$$

We use the notation $\partial_a x \equiv \partial x/\partial\lambda^a$ and subsequently will use $\partial_{ab} x \equiv \partial^2 x/\partial\lambda^a\partial\lambda^b$. If we now assume that the approximation is good, we can write $\delta h(\vec{\lambda}) \sim \mathcal{O}(\epsilon)$, a small parameter, and $\vec{\lambda}_{\rm bf} = \vec{\lambda}_0 + \Delta\vec{\lambda}$ with $\Delta\lambda^i \sim \mathcal{O}(\epsilon)\ \forall i$. We can then expand the difference between the approximate waveforms as a Taylor series

$$H(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_0) = \partial_a H(\vec{\lambda}_0)\Delta\lambda^a + \frac{1}{2}\partial_{ab} H(\vec{\lambda}_0)\Delta\lambda^a\Delta\lambda^b + \cdots\,. \tag{A3}$$

Eq. (A2) becomes

$$\left\langle \delta h(\vec{\lambda}_0) + \partial_b H(\vec{\lambda}_0)\Delta\lambda^b + \frac{1}{2}\partial_{bc} H(\vec{\lambda}_0)\Delta\lambda^b\Delta\lambda^c \middle| \partial_a H(\vec{\lambda}_0) + \partial_{ad} H(\vec{\lambda}_0)\Delta\lambda^d \right\rangle = 0\,. \tag{A4}$$

where all derivatives are now evaluated at $\vec{\lambda}_0$. Keeping only terms of order ϵ we find the Cutler and Vallisneri result

$$\Delta\lambda_1^a = -\left(\Sigma^{-1}\right)^{ab} \langle \delta h(\vec{\lambda}_0) | \partial_b H(\vec{\lambda}_0) \rangle \tag{A5}$$

where $\Sigma^{ij} \equiv \langle \partial_a H(\vec{\lambda}_0) | \partial_b H(\vec{\lambda}_0) \rangle$ is the Fisher Matrix.

We now extend to the next order in ϵ by writing $\Delta\lambda^a = \Delta\lambda_1^a + \Delta\lambda_2^a$, where $\Delta\lambda_1^i$ is the previous solution, Eq. (A5). Keeping terms to $\mathcal{O}(\epsilon^2)$ we obtain

$$\Delta\lambda_2^a = -\left(\Sigma^{-1}\right)^{ab} \left[\langle \delta h(\vec{\lambda}_0) | \partial_{ab} H(\vec{\lambda}_0) \rangle \Delta\lambda_1^b + \langle \partial_{ac} H(\vec{\lambda}_0) | \partial_b H(\vec{\lambda}_0) \rangle \Delta\lambda_1^b\Delta\lambda_1^c + \frac{1}{2} \langle \partial_a H(\vec{\lambda}_0) | \partial_{ab} H(\vec{\lambda}_0) \rangle \Delta\lambda_1^b\Delta\lambda_1^c \right]. \tag{A6}$$

A suitable validity criterion for the Cutler and Vallisneri formula, (A5), is

$$\max_a \{ |\Delta\lambda_2^a / \Delta\lambda_1^a| \} \ll 1. \tag{A7}$$

We note also that Eq. (A6) provides an improved estimate of the systematic bias and that we can readily extend this method to higher order in ϵ by including further terms in the expansion in Eq. (A3).